DOCUMENT RESUME

ED 480 441 CE 085 446

AUTHOR	Condelli, Larry; Baker, Holly
TITLE	Developing Performance Assessments for Adult Literacy Learners: A Summary.
INSTITUTION	National Academy of Sciences - National Research Council, Washington, DC. Board on Testing and Assessment.
SPONS AGENCY	Office of Vocational and Adult Education (ED), Washington, DC.
PUB DATE	2002-10-00
NOTE	14p.; Paper written for the project to "Improve the Quality and Use of National Reporting System (NRS) Data".
AVAILABLE FROM	For full text: http://www.nrsweb.org/reports/PerformanceAssessments.pdf.
PUB TYPE	Guides - Non-Classroom (055) -- Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; Accountability; Achievement Gains; *Adult Basic Education; Adult Literacy; *Evaluation Methods; Federal State Relationship; Field Tests; Guidelines; *Performance Based Assessment; Pilot Projects; Psychometrics; Scoring Rubrics; Standardized Tests; *State Standards; *Student Evaluation; *Test Construction; Test Format; Test Reliability; Test Validity
IDENTIFIERS	National Reporting System for Adult Education

ABSTRACT

This paper summarizes the standardized procedures and psychometric standards states must follow in developing performance-based assessment. The National Reporting System (NRS), which serves as the accountability system for federally funded adult education and literacy programs, requires that states have their local programs assess educational gain by selecting the assessments either published standardized tests or performance-based assessments most appropriate for their state. Performance-based assessments obtain richer information about students than do selected-response assessments, or standardized tests, and are often seen as more authentic because they resemble-real world tasks. But they cover less content, make it hard to generalize from one test to another, and are difficult and costly to score. In order to ensure that test scores resulting from performance-based assessments are valid and reliable according to the Standards for Educational and Psychological Testing, there are three steps in development. They are the following: (1) develop the tasks the test-takers will perform; (2) develop the scoring rubrics that will measure how well the test-takers performed on the tasks; and (3) pilot and field-test the assessment and scoring to determine the validity and reliability of the test items and scoring. Includes 4 references. (MO)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 480 441

Developing Performance Assessments for Adult Literacy Learners: A Summary

National Reporting System
for Adult Education

Division of Adult Education and Literacy
Office of Vocational and Adult Education
U.S. Department of Education

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

October 2002

CE 080 446

BEST COPY AVAILABLE

Developing Performance Assessments for Adult Literacy Learners: A Summary

Preface

The National Reporting System (NRS) *Implementation Guidelines* allow states to measure educational gain by pre- and posttesting students using a standardized assessment. States may use selected-response type assessments, such as multiple choice tests, or performance–based assessments, if they meet accepted psychometric standards, including having standardized administration procedures and scoring rubrics. Several states interested in developing performance based assessments have requested that the Office of Vocational and Adult Education (OVAE) clarify more explicitly the procedures and psychometric standards they must follow in developing assessments. In response, OVAE asked the Board on Testing and Assessment (BOTA) of the National Research Council of the National Academy of Sciences, to convene a panel of experts in assessment and adult education to provide guidance. The Board issued a report of the proceedings from the conference held by the panel, which included guidance on standardizing performance based assessments for the NRS.

This paper summarizes the guidance from the report on the process for developing a valid and reliable performance-based assessment. States that develop their own assessments must follow the procedures outlined is this paper to use the assessment for NRS accountability. The information in the paper should also be used to guide decisions on whether of existing assessments are valid and reliable and thus acceptable for use with the NRS.

Future revisions to the *NRS Guidelines* will incorporate this summary of the assessment development process. For more detailed explanations of these procedures, states should consult *Performance Assessments for Adult Education: Exploring the Measurement Issues* and *Standards for Educational and Psychological Testing*, referenced at the end of this paper.

This paper was written through the project to *Improve the Quality and Use of National Reporting System (NRS) Data*, U.S. Department of Education, Office of Vocational and Adult Education, Division of Adult Education and Literacy.

Developing Performance Assessments for Adult Literacy Learners: A Summary

Larry Condelli and Holly Baker
American Institutes for Research
Washington, DC

The National Reporting System (NRS) serves as the accountability system for the federally funded adult education and literacy program, under the authority of Title II of The Workforce Investment Act. The NRS identifies five core student outcome measures that states are to report to meet their accountability requirements under the Act, along with definitions of these measures, methodologies for collecting them and reporting formats. First published in 1999, the *NRS Implementation Guidelines* (March 2001) and *Guide for Improving NRS Data Quality* (August 2002) describe the required measures and procedures.

Educational gain, a key core outcome measure in the NRS, describes students' improvement in literacy skills during instruction. States are required to have their local programs assess gain by administering standardized pre-post assessments to students, following valid administration procedures (e.g., use an appropriate assessment, use different forms of the test for pre- and posttesting). The *NRS Guidelines* allows states to select the assessments most appropriate for their state, which may be published standardized tests or performance-based assessments. If the state uses performance-based assessments, NRS guidelines require the assessment to have standardized procedures and scoring rubrics that meet accepted psychometric standards.

Several states and programs have requested that the Office of Vocational and Adult Education (OVAE) clarify more explicitly the standardized procedures and psychometric standards they must follow in developing performance-based assessment. In response, OVAE asked the Board on Testing and Assessment (BOTA) of the National Research Council of the National Academy of Sciences, to convene a panel of experts in assessment and adult education to provide guidance. The BOTA's report, *Performance Assessments for Adult Education: Exploring the Measurement Issues* (May 2002) describe the proceedings from the conference held by the panel, which includes guidance on standardizing performance based assessments for the NRS.

This paper summarizes the steps required to develop a standardized performance assessment, based on generally recognized procedures for developing such assessments.

The summary draws from the BOTA's report, as well as the guiding document for assessment development in the field, *Standards for Educational and Psychological Testing* (1999).

Performance-Based Assessment

Assessment has three purposes: it provides diagnostic information (formative assessments), evaluates student progress (summative assessment) and evaluates the overall performance of an entity (e.g., class, program, state). Most assessment tools use two types of questions or tasks: selected-response items and constructed-response items, which include performance-based assessments.

The most familiar type of selected-response assessment is the multiple-choice test. These assessments offer several advantages. Many questions can be administered in a short period, which means that many content areas can be covered. Costs are fixed and scoring is fast and straightforward. Score reliability is high because the scoring process is objective. But selected-response assessments have some limitations. They often assess recall and recognition, not higher-order thinking and they are susceptible to guessing by test-takers. Because selected-response assessments do not require test-takers to apply their skills and knowledge, these tests do not provide information about students' responses to real-life situations.

These limitations, among others, have fueled a growing interest in performance assessments that require test-takers to demonstrate their skills and knowledge in a manner that closely resembles a real-life situation or setting.

Developing Performance-Based Assessments

Performance-based assessments have several advantages over selected-response assessments. They obtain richer information about students and they are often seen as more authentic because they resemble real-world tasks. But they have their own disadvantages, since they require more time to administer and use fewer items. As a result, performance-based assessments cover less content. Using fewer tasks makes it hard to generalize from one test to another. Scoring is difficult and costly because rubrics, or scoring guides, must be developed; scorers need extensive training; and scoring takes longer. These issues are serious for any assessment, but are especially crucial in high-stakes circumstances.

Reliability and Validity

Like all other assessments, the value of the information provided by the performance assessments depends on the reliability and validity of the measures. The *Standards for Educational and Psychological Testing* defines *reliability* as "the consistency of … measurements when the testing procedure is repeated on a population of individuals or groups." In essence, differences in items, scorers, administration procedures and setting do not affect the consistency of a reliable assessment.

The *Standards* define *validity* as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" and therefore focuses on the scores, their interpretation and their uses, not on the assessment itself. According to the *Standards*, validity also depends on the fairness of the assessment, the comparability of different versions of the assessment and its generalizability. Fairness has four aspects: lack of bias, equitable treatment in the testing process, equality in outcomes of testing and opportunity to learn. Comparability means that a test has different versions that "yield scores that can be used interchangeably even though they are based on different sets of items." Generalizability occurs when test results are applied to a population beyond that of the original test-takers and to an entire knowledge domain beyond the areas sampled by the test.

Development Process

The development of performance assessments is essentially a process to ensure that the resulting test scores are valid and reliable according to these standards. There are three steps to developing performance-based assessments:

- **Develop the tasks** that test-takers will perform;
- **Develop the scoring rubrics**, which measures how well the test-taker performed on the tasks; and
- **Pilot and field-test** the assessment and scoring, to determine the validity and reliability of the test items and scoring.

Once field-testing is completed, the final assessments and its various versions or forms are assembled.

Exhibit 1 summarizes the development process for performance-based assessments, which we describe in more detail below. Although we focus on performance assessment, the psychometric issues and processes discussed are the same as those for any type of assessment.

Developing the Tasks

Developing the tasks for the assessment first entails defining the population of test-takers, the purpose of the assessment and the *domain* to be covered, which includes the content and skills to be assessed (e.g., basic reading skills, reading comprehension, grammar, writing) and the number and types of items assessing a particular skill.

The tasks of the assessment must reflect the underlying skills desired and must be appropriate for the population of test-takers. For assessments designed to measure the relationship of instruction to student outcomes, such as for NRS accountability or to evaluate local program performance, the skills measured by the assessment should match the curriculum of instruction. For example, if a state is developing performance-based

Exhibit 1
Summary of Performance Assessment Development Process

Step 1: Develop the Tasks

- Define the domain to be covered by identifying the population of test-takers, the purpose of the assessment and the content and skills to be assessed.
- Develop and evaluate test specifications that identify the content of the assessment and the number and types of items assessing each skill. Tie assessment to state instructional content standards if it is to be used for NRS accountability
- Develop the actual items or tasks and review them for content, clarity, lack of ambiguity, sensitivity to gender or cultural issues, and fairness.
- Develop the response formats — how the learner is to perform the task or answer the item.

Step 2: Develop Scoring Rubrics

- Select the type of rubric: generic or task specific, and analytic or holistic; and the scaling of the rubric.
- Decide on the skills the rubrics will measure and the scoring categories. Categories should have about equal scoring increments among levels.
- Train scorers and conduct field studies to establish the reliability and accuracy of scoring among raters.

Step 3: Pilot and Field Test the Assessment

- Conduct cognitive lab studies to provide information on how test-takers approach and understand the assessment.
- Conduct pilot tests and small-scale tryouts to evaluate assessment administration procedures, tasks and scoring rubrics.
- Conduct large-scale field tests to again evaluate administration and scoring, and to obtain data on the statistical properties of the tests.
- Compute reliability statistics for scoring rubrics.
- Assemble final assessment, selecting the most valid items and tasks and reliable scoring procedures.
- Document administration, scoring and the assessment's statistical properties.

assessments in a domain, the state should first delineate the instructional content standards for the domain, which will guide both instructional content and the assessment.

Performance assessments have the potential to supply deep and rich information about test-takers. But because they are time intensive, they can use only a few tasks. Therefore the limitation of these assessments to represent a wide scope of content and skills is a significant issue. That is, how many items are necessary to adequately assess the skill? Two approaches for resolving this problem – technically known as achieving domain coverage – are the critical indicator approach and the domain sampling approach.

In the critical indicator approach, specific skills in a given content domain are singled out as being more important than others. This approach is based on the assumption that a test-taker's performance on the critical task represents his or her mastery of the entire content domain. The approach has two limitations. First, results are based on a small sample. Second, test developers must agree on the domain, understand the skills and knowledge that the domain requires and select the critical tasks.

In the domain sampling approach, a large pool of items is developed that represent all the essential skills and tasks in the domain. To construct a test, developers randomly pull items from this pool. This approach is based on the assumption that a test-taker's performance on the sample of items represents his or her mastery of the entire content domain. Like developers who use the critical indicator approach, developers who use domain sampling must agree on the domain and understand the skills and knowledge that the domain requires. Additional limitations are the amount of time that is needed to ensure good domain coverage and the lack of agreement about what the domain should cover.

Once the domain has been delineated, the next step is to develop the actual items or tasks. This process includes both writing items and reviewing them for content, clarity, lack of ambiguity, sensitivity to gender or cultural issues and fairness. Performance-based assessment encompasses a number of types of tasks. The three most prominent are performance tasks, written scenarios and portfolios.

Responses to Tasks

The task development process includes response specifications — how the learner is to perform the task or answer the item. With performance-based assessment, test-takers must usually perform what is often called an "authentic" task. These tasks are designed to show test-takers' knowledge and skills in a real-world context. For example, a test-taker might be asked to follow the steps necessary to apply for a job. The tasks could include reading the want ads, using a map to plot a route to the employer and filling out an application.

When the assessment requires a written scenario, the test-taker is given a writing prompt that requires him or her to use knowledge and skills to write a solution to a real-world problem. The test-taker is given the scenario, writing instructions and evaluation

criteria. These assessments are easy to develop, modify and administer, but they are difficult to score.

To develop a portfolio, students gather examples of their work that they believe represent their best efforts in different domains. Often specific categories of examples are required. For accountability purposes, portfolios should also include work samples from tasks completed when the student first begins class to allow benchmarks of initial abilities. The initial performance on tasks can then be compared to subsequent work. Although portfolios can be an excellent way to track the academic growth of an individual student, their idiosyncratic nature makes them less valuable as a way to measure the effectiveness of large entities or programs.

In determining responses, the test developer must consider not only whether the responses accurately captures the domain, but the time it takes for the test-taker to produce the responses. A portfolio, for example, may take several weeks or months for a student to produce. Scoring of the responses is also a critical consideration in designing the task, since the more complex the task, the longer it will take to score and the more difficult it will be to develop reliable scoring.

Develop Scoring Rubrics

Rubrics, or scoring guides, describe the key features that a response must include to merit a specific score. In many cases, writing the rubrics is more difficult than developing the tasks. Good rubrics must be carefully and unambiguously defined to ensure a reliable score. The reliability of scores is absolutely essential to the quality of the assessment.

Selecting the Type of Rubric

Although all rubrics are scoring guidelines for judging the correctness of a test-taker's response to an item, different types of rubrics are available. Rubric writers must make several decisions before beginning rubric development.

Writers must first decide whether a generic rubric or a task-specific rubric will work best for a given task or for an overall assessment. A generic rubric can be applied to a wide variety of tasks because it identifies underlying cognitive abilities. A task-specific rubric is relevant to only a specific task; it is customized to elicit very specific knowledge and skills. For example, a generic rubric could be used to evaluate a test-taker's writing ability. The corresponding task-specific rubric could be designed to evaluate his or her ability to write a declarative sentence. Generic rubrics are more useful because of their flexibility, whereas task-specific rubrics give more details about a test-taker's performance.

The next decision is a choice between analytic and holistic rubrics. Analytic rubrics separate out different aspects of a task and give a distinct score to each one. Holistic rubrics score a task as a whole and use only one scale, which encompasses all

aspects of the task. Analytic rubrics give more detail; holistic rubrics are more efficient to develop and use.

The third preliminary decision is the type of scale to use. Some scales are qualitative and give a performance a label, such as below average, average and above average. A quantitative scale assigns a number of points to different levels of performance, such as 0 for a completely incorrect response to 4 for a fully correct response. Many rubrics use a combination of these approaches.

Developing the Rubrics

The first steps in writing a good scoring rubric for a performance assessment are the same as those used to write the task: deciding on the content and skills to be assessed and choosing the appropriate task to measure these skills and content. The next step is deciding on the type of rubric. The remaining steps are common to all types of assessments.

Rubric writers must decide on the range of acceptable performance for the task. What skills and content knowledge must a test-taker demonstrate as evidence that he or she has reached mastery? What would the best performance look like? The worst performance? How good is good enough? Rubric writers then break these expectations into measurable details. Once the details are delineated, the writers decide the number of score levels that will best separate the lowest and highest anticipated performances.

Some rubrics have just two levels: the response is correct or incorrect, with no possible gradations. Others have multiple levels. However, no ideal number of levels exists. Too few may not adequately distinguish among responses; too many can be unwieldy for the scorers to use. The differences between degrees of response correctness must represent meaningful, not trivial, distinctions in test-takers' achievement. In practice, once the writers choose what the top and bottom responses should include, the number of levels in the middle becomes clear.

A good rubric has one essential feature, regardless of the number of score levels: it clearly distinguishes one level from the next. For precision, the rubric should use words that describe specific work; it should not use vague judgmental terms. For example, "The student properly used standard punctuation marks in declarative sentences" is better than "The student has good punctuation skills."

The rubric should also require approximately equal increments of performance from one score level to the next. In other words, the degrees of response correctness require roughly equal jumps in achievement; on a 0 to 5 scale, the rubric should not demand a significantly greater leap in achievement to go from score level 2 to level 3 than to go from level 4 to level 5. No matter how clear and well written the rubrics are, embedding actual examples of a potential response will greatly help scorers.

One part of the rubric that is often overlooked and under appreciated is the set of directions for the test-takers. A good rubric tells them on what they are being assessed and how the scoring will be done. For example, if the top score requires test-takers to give five examples of something, the rubric should say so in explicit and jargon-free language.

Training Scorers

Although the rubric itself is important, the scorers who use the rubric are critical to the success of a performance-based assessment. Since scorers are only human, they have prejudices and preferences, get tired or bored and can be influenced by appearance over substance. A good scoring process eliminates this scorer variability as much as possible by keeping the scorers on track and working to a common standard. To achieve this, scorers must be extensively and carefully trained to ensure that scoring is consistent across scorers and across tasks. That is, the same performance-based item should receive the same score regardless of who the scorer is — the essence of performance assessment reliability. Scoring procedures are evaluated and refined as part of the pilot testing of the assessment.

Scorers are ideally trained to use each rubric. They review the rubric, discuss examples of responses that have been previously scored by experts and practice scoring a pre-selected set of responses. Once scorers have passed through training, they score assessments from field and pilot-tests of the assessment where their work is subject to intense quality checks. For example, they might be given a response that an expert scorer has already scored or score the same responses again. In both cases, the scores must agree. Throughout the scoring activity, scoring experts who are also content specialists may randomly select a scored response to check on the quality of the scoring.

Pilot and Field Testing

After the tasks, test-taker responses and scoring rubrics have been developed, the assessment developers review the tasks, responses and scoring along the following dimensions.

- *Match of items and tasks to the domain specifications.* This review ensures that items are consistent with the domain being assessed.
- *Scorability review.* This review ensures that the scoring rubrics are directly relevant to the tasks items and are likely to elicit responses that can be scored reliably and validity.
- *Bias and sensitivity/language simplification.* This review ensures that tasks and items (a) avoid issues and language that may reflect biases (e.g., race/ethnic stereotyping), (b) avoid topics that touch on sensitive issues (e.g., religion), (c) reflect the diversity of the examinee population and (d) are written in clear language that minimizes language-based complexities that are not directly related to the constructs being assessed.

Test developers often use an external committee of assessment and content experts to assist in the review.

Next, pilot and field tests of the assessment are conducted. The purpose of the pilot tests is to try the assessment on samples of the test-taking population to determine whether that procedures related to the administration and scoring of the test are practical and reliable. In addition, the field tests provide data to determine the psychometric properties of the test through statistical analyses.

Most assessments receive three types of pilot studies: cognitive lab procedures, small-scale pilots and tryouts, and much larger field tests. Cognitive labs provide information on how test-takers approach and understand the assessment, while pilot tests offer test developers an early indication of how the test and scoring will work with small samples. Field tests are usually larger scale studies that also provide information on the statistical properties of the tests. The extent of field testing and documenting of statistical properties of a test required to ensure a valid and reliable assessment depends partly on the planned use of the test. Assessments used for high stakes decision-making, such as for accountability and funding, require greater levels of statistical evidence of validity and reliability, and thus need more intensive field testing and psychometric documentation.

Cognitive Labs

A cognitive lab is a small-scale study where an interviewer administers the assessment individually to a test-taker and then guides the test-taker through a retrospective explanation of how he or she determined the answers. Through cognitive lab studies, test developers collect in-depth qualitative information on assessment instruments to answer questions such as whether the directions are clear, whether test-takers understand the tasks and items as intended by the developers and whether there is enough time to complete the assessment tasks. Cognitive lab research also investigates whether test administrators understand what they need to do, what they do that may help or hinder performance, whether directions for scoring the assessment are clear and if the wording of scales and rubrics is clear. Since cognitive labs are expensive, they are usually conducted on only about 10 – 20 test-takers.

Small Scale Pilot and Tryout

Small-scale pilot and tryout studies employ small samples of test-takers from the target population. Depending on the complexity, planned use and importance of the assessment for high stakes decision-making, several pilots and tryouts may be conducted, each with a minimum sample size of about 30 test-takers. The pilots provide information on the clarity of the directions to test administrators, the clarity of the directions to test-takers and their performance in completion of assessment tasks. They also provide information on the scoring rubrics, including how well they work and the training scorers need to ensure reliability and validity. Assessment developers use data from pilot and tryout studies to revise items and test administration procedures as necessary.

Field Tests

Using the revised assessment, test developers next conduct a larger field test of the assessment. The field test is designed to examine how the assessment works under normal operational testing conditions with a large sample of test-takers. Again, depending on the complexity, planned use and whether the assessment will be used for high stakes decisions, several field tests may be conducted, with sample sizes ranging from a minimum of 300 to several thousand test-takers. High stakes assessments have the more rigorous requirements.

Not only does the field test provide information on procedures for administration and reliability of scoring of the assessment, but data are used to determine whether tasks and items match the assessment domains and framework and to develop statistics on the validity and reliability of the test. If the field test is large enough, data can also be used to provide norms for the test and to explore fully its psychometric properties.

Psychometricians use several complex statistic methods, including item response theory (IRT), to identify which tasks measure the assessment's constructs, the number of items the assessment needs for a given level of validity and reliability, the difficulty of the tasks and items and the reliability of scoring. These data may suggest further revisions to the assessment and scoring rubrics, which in turn require additional field-testing. Scorer agreement is also quantitatively determined from field-testing through reliability analyses. A good assessment typically requires greater than 80 percent reliability. In high stakes assessment, reliability greater than 90 percent is often required.

Assessment Assembly

Once all field-testing is completed, the test developers use the results of the various pilot and field tests to create the actual assessment. Factors considered in designing the final assessment include the item statistics on validity and reliability, the scoring properties and training needed for scorers to achieve a high level of reliability, and the time to administer and score the test. Test developers select the actual tasks or items for the assessment and arrange them in a specific sequence to create a version or form of the assessment. If the assessment is to be used for pre- and posttesting, multiple forms of the assessment must be developed.

Along with the assessment, test developers usually publish a technical manual on the test that includes reliability and validity statistics and norming information, if any. For performance assessment a detailed manual on scoring is provided, as well as a description of the standardized administration procedures. To implement the assessment, states need to train staff on using the standardized procedures and scoring the assessment to obtain high levels of scoring reliability.

References

American Educational Research Association (1999). *Standards for Educational and Psychological Testing.* Washington, D.C.

National Academy of Sciences (2002). *Performance Assessments for Adult Education: Exploring the Measurement Issues.* National Research Council, Board on Testing and Assessment, Washington, D.C.

U.S. Department of Education (2001). *Measures and Methods for the National Reporting System for Adult Education: Implementation Guidelines.* Office of Vocational and Adult Education, Division of Adult Education and Literacy, Washington, D.C.

U.S. Department of Education (2002). *Guide for Improving NRS Data Quality: Procedures for Data Collection and Training.* Office of Vocational and Adult Education, Division of Adult Education and Literacy, Washington, D.C.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[ ] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[X] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").